

Scalable Steiner Tree for Multicast Communications in Software-Defined Networking

Liang-Hao Huang, Hui-Ju Hung, Chih-Chung Lin, and De-Nian Yang

Academia Sinica, Taipei, Taiwan

{lhhuang, hjhung, chchlin, dnyang}@iis.sinica.edu.tw

Abstract—Software-Defined Networking (SDN) enables flexible network resource allocations for traffic engineering, but at the same time the scalability problem becomes more serious since traffic is more difficult to be aggregated. Those crucial issues in SDN have been studied for unicast but have not been explored for multicast traffic, and addressing those issues for multicast is more challenging since the identities and the number of members in a multicast group can be arbitrary. In this paper, therefore, we propose a new multicast tree for SDN, named Branch-aware Steiner Tree (BST). The BST problem is difficult since it needs to jointly minimize the numbers of the edges and the branch nodes in a tree, and we prove that it is NP-Hard and inapproximable within k , which denotes the number of group members. We further design an approximation algorithm, called Branch Aware Edge Reduction Algorithm (BAERA), to solve the problem. Simulation results demonstrate that the trees obtained by BAERA are more bandwidth-efficient and scalable than the shortest-path trees and traditional Steiner trees. Most importantly, BAERA is computation-efficient to be deployed in SDN since it can generate a tree on massive networks in small time.

Index Terms—SDN, multicast, NP-Hard, traffic engineering, scalability

I. INTRODUCTION

Software-Defined Networking (SDN) is an emerging architecture that is manageable, dynamic, cost-effective, and adaptable, making it ideal for the high-bandwidth, huge data, and dynamic nature of numerous network services [1]. This novel architecture decouples the network control and forwarding functions. It enables the network control to become directly programmable and the underlying infrastructure to be abstracted for varied applications. The OpenFlow protocol has been recognized as a crucial element for building SDN solutions [1], [2], [3].

SDN comprises two main components: SDN controller (SDN-C) and SDN forwarding element (SDN-FE) [3]. Compared with the traditional shortest-path routing, SDN-C enables the centralized computation on unicast routing for traffic engineering [4] to improve the network throughput. Nevertheless, since the routing paths no longer need to be the shortest ones, the paths can be distributed flexibly inside the network and thus are more difficult to be aggregated in the flow table of SDN-FE, and the scalability has been regarded as a serious issue to deploy SDN in a large network [4], [5].

Multicast is an efficient technique for point-to-multipoint (P2M) and multipoint-to-multipoint (M2M) communications because it exploits a tree, instead of disjoint paths, in the routing of the traffic. Current multicast standard on Internet, i.e., PIM-SM [6], employs a shortest-path tree to connect the

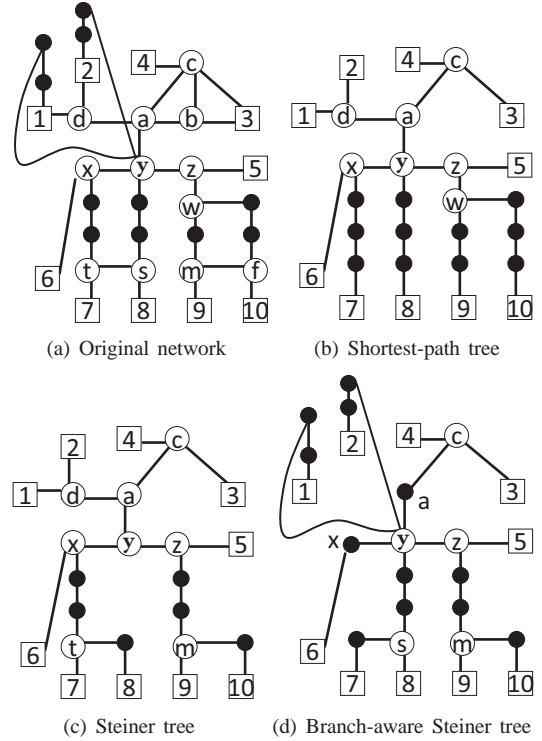


Fig. 1. An example of multicast tree

terminal nodes in a multicast group, where a terminal node is a designated router connecting to a LAN with at least one user client joining the group [7]. Traffic engineering is difficult to be supported in a shortest-path tree since the path from the root, i.e., the traffic source in P2M or the rendezvous point in M2M in PIM-SM, to each destination in the tree is still the shortest path. By contrast, a Steiner Tree (ST) [8] in Graph Theory is more promising because it minimizes the network resource consumption, i.e., the number of edges in a tree, required for a multicast group. However, finding an ST is more computation intensive and thus is difficult to be deployed as a distributed protocol on Internet. By contrast, now it becomes feasible by first finding an ST in SDN-C and then storing the forwarding information in the group tables of SDN-FEs on the tree.

Similar to unicast traffic engineering in SDN, multicast traffic engineering also suffers from the scalability problem since each SDN-FE in the tree needs to store a forwarding entry in the group table for each multicast group. Nevertheless, the scalability problem for multicast communications

is even more serious since the number of possible multicast group is $O(2^n)$, where n is the number of nodes in a network, and the number of possible unicast connections is $O(n^2)$. To remedy this issue, a promising way is to exploit the *branch forwarding technique* [9], [10], [11], [12], [13], which stores the entries in only the branch nodes, instead of every node, of a multicast tree. More specifically, a branch node in a tree is the node with at least three incident edges, such as white circle nodes in Fig. 1, and the square nodes are the terminal nodes. To minimize the total number of edges in an ST, the path connecting two neighboring branch nodes (such as nodes c and y in Fig. 1(d)) needs to be the shortest path between them. Note that an ST is not a shortest-path tree because the branch nodes can be located anywhere in the network. This branch forwarding technique can remedy the multicast scalability problem since packets are forwarded in a unicast tunnel from the logic port of a branch node in SDN-FE [3]. In other words, all nodes in the path (such as black circle nodes in Fig. 1) exploit unicast forwarding in the tunnel and are no longer necessary to maintain a forwarding entry for the multicast group.

To effectively address the multicast scalability problem in SDN, it is crucial to minimize the number of branch nodes in a tree. However, this important factor has not been considered in ST. In this paper, therefore, we propose a new multicast tree for SDN, named *Branch-aware Steiner Tree (BST)*. The objective of BST problem is to minimize the summation of the number of edges and the number of branch nodes in the tree, where a branch node can be assigned a higher weight to further improve the scalability. Fig. 1 presents an illustrative example with the weight of each branch node set as 20. Square nodes are the terminal nodes that are required to be connected in a tree, while the black and white circle nodes are the other nodes in the network. Fig. 1(a) is the network topology. The shortest-path tree in Fig. 1(b) includes 27 edges and 7 branch nodes with the total cost of the tree as $27 + 7 \times 20 = 167$. The Steiner tree in Fig. 1(c) has 23 edges and 8 branch nodes with the total cost as $23 + 8 \times 20 = 183$. By contrast, Fig. 1(d) presents the BST with 26 edges and 5 branch nodes and the total cost as $26 + 5 \times 20 = 126$. Therefore, compared with the shortest-path trees on Internet, BST effectively reduces the network resource consumption by minimizing the number of edges in the tree. Compared with ST, more BSTs can be supported in SDN since the number of branch nodes is effectively minimized.

Finding an BST is very challenging. The ST problem is NP-Hard but can be approximated within ratio 1.55 [14] and is thus in APX of complexity theory. In other words, there exists an approximation algorithm for ST that can find a tree with the total cost at most 1.55 times of the optimal solution. By contrast, we prove that BST is NP-Hard but cannot be approximated within k , which denotes the number of terminal nodes in a multicast group. In other words, the BST problem is more difficult to be approximated. To effectively solve BST, we propose a k -approximation algorithm, named *Branch Aware Edge Reduction Algorithm (BAERA)*, that can be deployed in SDN-C. BAERA includes two phases, Edge Optimization Phase and Branch Optimization Phase,

to effectively minimize the number of edges and branch nodes. Since no $(k^{1-\epsilon})$ -approximation algorithm exists in BST for arbitrarily small $\epsilon > 0$, BAERA achieves the best approximation ratio.

The rest of this paper is organized as follows. Section II briefly summarizes the literature on SDN traffic engineering, SDN flow table scalability, multicast scalability, and the Steiner tree. Section III formally presents the problem formulation with Integer Programming and the hardness result. We design a k -approximation algorithm in Section IV, and Section V presents the simulation results to evaluate the performance of the proposed algorithm in real networks. We conclude this paper in Section VI.

II. RELATED WORKS

Previous works have extensively explored the issues on traffic engineering and flow table scalability for *unicast traffic* in SDN. Mckeown et al. [2] pointed out that OpenFlow can be deployed with heterogeneous switches. Sushant et al. [15] shared their experience of SDN development for the private WAN of Google Inc. Qazi et al. [16] proposed a new system design using SDN for the middleboxes (e.g., firewalls, VPN gateways, proxies). Agarwal et al. [4] considered the incremental deployment of traffic engineering in the case where a SDN-C controls only a few SDN-FEs in the network, and the rest of the network adopts a standard routing protocol, such as OSPF. The merits of traffic engineering brought by only a limited number of SDN-capable nodes are demonstrated. Mueller et al. [17] presented a cross-layer framework in SDN, which integrates a novel dynamic traffic engineering approach with an adaptive network management, to bridge the gap between the network and application layers for overall system optimizations.

On the other hand, flow table scalability is crucial to enable a large-scale deployment of SDN. For unicast traffic, Kanizo et al. [5] pointed out that the restriction on table sizes is the major bottleneck in SDN and proposed a framework, called Palette, to decompose a large SDN table into small ones and then distribute them across the network. Lee et al. [18] observed that Data Center traffic frequently meets few elephant flows and a lot of mice flows. However, elephant flows are inclined to be evicted because of the limited flow table sizes. They proposed a differential flow cache framework that uses a hash-based cache placement and localized Least Recently Used (LRU)-based replacement to reduce the loss of elephant flows.

The scalability issue is more serious in multicast, and the previous works [9], [10], [11], [12], [13] have demonstrated that the branch forwarding technique is a promising way since forwarding from a branch node to a neighbor branch node or terminal node can exploit the existing unicast tunneling technique, and tunneling can be facilitated in SDN with logic ports specified in the group table [3]. In other words, the intermediate nodes between two neighbor branch routers no longer need to store a multicast forwarding entry for the tree. However, the above works were designed for shortest-path trees and did not explore the possibility of more flexible multicast routing. On the other hand, Steiner tree [8] can effectively minimize the bandwidth consumption in a network, but so far it is not adopted on Internet since finding

the optimal Steiner tree is more computation intensive and thus difficult to be deployed as a distributed protocol. To remedy this issue, overlay Steiner trees [19], [20] for P2P environments are proposed, where only the terminal nodes can act as branch nodes. Nevertheless, the merit of traffic engineering from the above work is limited since no other router can act as the branch node to reduce the bandwidth consumption. Moreover, multicast scalability is not studied in the above works. Therefore, the above works are difficult for bandwidth-efficient and scalable multicast in SDN.

III. PRELIMINARIES

A. Problem Formulation

In this paper, we propose a scalable and bandwidth-efficient multicast tree for SDN, called *Branch-aware Steiner Tree (BST)*. This paper aims to minimize the bandwidth consumption (i.e., the total number of links/edges) and the number of forwarding entries maintained for the multicast group (i.e., the total branch nodes). Therefore, the BST problem is to find a tree connecting a given set of terminal nodes such that the sum of the number of edges and the number of branch nodes is minimized, where a branch node can be assigned a larger weight w to ensure a higher scalability.¹

Definition 1: Consider a network $G(V, E)$, where V and E denote the set of nodes and edges, respectively. Given $G(V, E)$, a terminal node set $K \subseteq V$, and a non-negative value w , the BST problem is to find a tree T spanning the terminal node set K such that $c(T) + b(T)w$ is minimized, where $c(T)$ is the number of edges on T , and $b(T)$ is the number of branch nodes (i.e., nodes with the degree at least 3 on T).

In BST, a network operator can increase the scalability of multicast in SDN by assigning a larger weight w for branch nodes. Compared with ST, $c(T)$ may slightly increase, but much fewer branch nodes will be selected in T . Compared with the shortest-path trees adopted on Internet currently, BST allows more flexible routing of a tree and thus can effectively reduce the network resource consumption and improve the scalability in SDN.

In the following, we first formulate the BST problem as an Integer Programming problem. Afterwards, we show that the BST problem is very challenging in complexity theory by proving that it is NP-Hard and not able to be approximated within k^c for every $c < 1$.

B. Integer Programming

Let N_v denote the set of neighbor nodes of v in G , and u is in N_v if $e_{u,v}$ is an edge from u to v in E . Let any terminal node r act as the root of T , i.e., the source, and the destination set L contains the other terminals in K , i.e., $L = K - \{r\}$. The output tree T needs to ensure that there is only one path in T from r to every node in L . To achieve this goal, our problem includes the following binary decision variables. Let binary variable $\pi_{l,u,v}$ denote if edge $e_{u,v}$ is

in the path from r to a destination node l in L . Let binary variable $\varepsilon_{u,v}$ denote if edge $e_{u,v}$ is in T , where $\varepsilon_{u,v} = \varepsilon_{v,u}$. Let binary variable β_v denote if v is a branch node in T . Intuitively, when we are able to find the path from r to each destination node l with $\pi_{l,u,v} = 1$ on every edge $e_{u,v}$ in the path, the routing of the tree with $\varepsilon_{u,v} = 1$ for every edge $e_{u,v}$ in T can be constructed with the union of the paths from r to all destination nodes in L , and every branch node v in T with $\beta_v = 1$ in T can be identified accordingly.

Most importantly, to guarantee that the union of the paths is a tree, i.e., a subgraph without any cycle, the objective function of our Integer Programming formulation (IP) is as follows.

$$\min \sum_{e_{u,v} \in E} \varepsilon_{u,v} + \sum_{v \in V} w \times \beta_v.$$

If the tree T contains any cycle, T is not optimal since we are able to remove at least one edge from the cycle to reduce the objective value, and ensure that there still exist a path from r to every destination node l in L . To find $\varepsilon_{u,v}$ and β_v from $\pi_{l,u,v}$, our IP formulation includes the following constraints.

$$\sum_{v \in N_r} \pi_{l,r,v} - \sum_{v \in N_r} \pi_{l,v,r} = 1, \forall l \in L, \quad (1)$$

$$\sum_{u \in N_l} \pi_{l,u,l} - \sum_{u \in N_l} \pi_{l,l,u} = 1, \forall l \in L, \quad (2)$$

$$\sum_{v \in N_u} \pi_{l,v,u} = \sum_{v \in N_u} \pi_{l,u,v}, \quad (3)$$

$$\forall l \in L, \forall u \in V, u \neq l, u \neq r, \quad (4)$$

$$\pi_{l,u,v} \leq \varepsilon_{u,v}, \forall l \in L, \forall e_{u,v} \in E, \quad (4)$$

$$\frac{1}{|N_u|} \left(-2 + \sum_{v \in N_u} \varepsilon_{u,v} \right) \leq \beta_u, \forall u \in V. \quad (5)$$

The first three constraints, i.e., (1), (2), and (3), are the flow-continuity constraints to find the path from r to every destination node l in L . More specifically, r is the flow source, i.e., the source of the path to every destination node l , and constraint (1) states that the net outgoing flow from r is one, implying that at least one edge $e_{r,v}$ from r to any neighbor node v needs to be selected with $\pi_{l,r,v} = 1$. Note that here decision variables $\pi_{l,r,v}$ and $\pi_{l,v,r}$ are two different variables because the flow is directed. On the other hand, every destination node l is the flow destination, and constraint (2) ensures that the net incoming flow to l is one, implying that at least one edge $e_{u,l}$ from any neighbor node u to l must be selected with $\pi_{l,u,l} = 1$. For every other node u , constraint (3) guarantees that u is either located in the path or not. If u is located in the path, both the incoming flow and outgoing flow for u are at least one, indicating that at least one binary variable $\pi_{l,v,u}$ is 1 for the incoming flow, and at least one binary variable $\pi_{l,u,v}$ is 1 for the outgoing flow. Otherwise, both $\pi_{l,v,u}$ and $\pi_{l,u,v}$ are 0. Note that the objective function will ensure that $\pi_{l,v,u} = 1$ for at most one neighbor node v to achieve the minimum cost. In other words, both the incoming flow and outgoing flow among u and v cannot exceed 1.

Constraints (4) and (5) are formulated to find the routing of the tree and its corresponding branch nodes, i.e., $\varepsilon_{u,v}$ and β_v . Constraint (4) states that $\varepsilon_{u,v}$ must be 1 if edge $e_{u,v}$ is included in the path from r to at least one l , i.e.,

¹Note that this problem can be simply extended to support different weights on each edges and each nodes. For example, a congested edge or a node with the group table almost full can be assigned a higher weight.

$\pi_{l,u,v} = 1$. The tree T is the union of the paths from r to all destination nodes. Note that here $\varepsilon_{u,v}$ and $\varepsilon_{v,u}$ represent the same binary decision variable because T is not directed. In other words, $\varepsilon_{u,v} = 1$ if edge $e_{u,v}$ is in a path (i.e., a directed flow) from either direction. The last constraint is the most crucial one. For each node u , if the degree of u is at least 3 in T , $\sum_{v \in N_u} \varepsilon_{u,v} \geq 3$ holds, and thus the left-hand-side of constraint (5) becomes positive, thereby enforcing that $\beta_u = 1$ and u acts as a branch node. Otherwise, the left-hand-side of constraint (5) is 0 or negative, allowing β_u to be 0 to minimize the cost in the objective function. In this case, node u is not a branch node in T .

C. Hardness result

The BST problem is NP-Hard because it is equivalent to the ST problem when w is 0. In other words, the ST problem is a special case of the BST problem. However, the BST is much more challenging because the ST problem can be approximated within ratio 1.55 [14] and is thus in APX in complexity theory, but we find out that BST is much more difficult to be approximated. The following theorem proves that the BST problem cannot be approximated within k^c for every $c < 1$, by a gap-introducing reduction from the Hamiltonian path problem, which determines whether there exists a path going through every node on a graph exactly once.

Theorem 1: For any $\epsilon > 0$, there exists no $k^{1-\epsilon}$ approximation algorithm for the BST problem, assuming $P \neq NP$.

Proof: We prove the theorem with the gap-introducing reduction from the Hamiltonian path problem. For an instance $G_H(V_H, E_H)$ of the Hamiltonian path problem with any node v on G_H , we build an instance of the BST problem on $G(V, E)$, such that

- if a Hamiltonian path starting at v exists in G_H , $\text{OPT}(G) \leq 2h$, and
- if no Hamiltonian path starting at v exists in G_H , $\text{OPT}(G) > 2hk^{1-\epsilon}$

, where h is the number of nodes in G and $\text{OPT}(G)$ is the optimal solution of G for the BST problem.

We first detail how to build the instance of the BST problem from the Hamiltonian path problem. For any given G_H , we construct a new graph G which consists of n^p copies of G_H , where n is the number of nodes in G_H and p is the smallest integer following $p \geq \frac{2}{\epsilon}$. One additional node x is added to G to connect to the node v of each of the n^p copies. The K is set to $V - \{x\}$ and w is set to h , where h is the number of nodes in G , i.e., $h = (n^p) \times n + 1$.

If G_H has a Hamiltonian path starting at v , consider a tree rooted at x , which includes 1) the edges between x and v of all copies and 2) the edges on the Hamiltonian path of all copies. The tree is a feasible solution of the BST problem with only one branch node x , and it can act as an upper bound of the BST in G . Thus, $\text{OPT}(G) \leq h + (h - 1) < 2h$. On the other hand, if G_H does not have a Hamiltonian path starting at v , there must exist at least one additional branch node in each copy of G . Hence, $\text{OPT}(G) > hn^p \geq 2hn^{p-1} = 2h(n^{p+1})^{\frac{p-1}{p+1}} \geq 2h(n^{p+1})^{1-\epsilon} = 2hk^{1-\epsilon}$. Since ϵ can be arbitrarily small, for any $\epsilon > 0$, there is no $k^{1-\epsilon}$ approximation algorithm for the BST problem, assuming $P \neq NP$. The theorem follows. ■

IV. ALGORITHM DESIGN

For BST, the shortest-path tree is not a good solution since the shortest path for each node v in K is constructed individually. With the aim to minimize the number of the edges, substituting the shortest path of v with a longer path can reduce the total edge number when the path mostly overlaps with the path to another node $v' \in K$ [8]. Therefore, it is expected that aggregating two paths that share more common edges can effectively reduce the number of edges in T . Nevertheless, aggregating two paths that partially overlap will generate a new branch node, and more branch nodes are inclined to be created when more paths are aggregated. Without considering the number of branch nodes created, the solution quality may deteriorate even though the number of edges in T is effectively reduced. In the following, therefore, we propose a k -approximation algorithm for BST, called *Branch Aware Edge Reduction Algorithm* (BAERA), to jointly minimize the numbers of edges and branch nodes in T . As Theorem 1 proves that no $(k^{1-\epsilon})$ -approximation algorithm for any $\epsilon > 0$ for the BST problem, BAERA achieves the best approximation ratio. Due to space constraint, the pseudo code is presented in [21].

BAERA includes two phases: 1) Edge Optimization Phase and 2) Branch Optimization Phase. In the first phase, BAERA iteratively chooses and adds a terminal node in K to the solution tree $T(V_T, E_T)$ for constructing a basic BST, where V_T and E_T denote the nodes and edges currently in T , respectively at each iteration. Initially, a random root node is added to V_T . Afterwards, for each terminal node $v \in K$ that is not in V_T , BAERA first finds the minimal distance $d_{v,T}$ from v to T . Precisely, let $p_{v,u}$ denote the shortest path from v to u on the network G , and $|p_{v,u}|$ is the number of edges in $p_{v,u}$. The minimal distance $d_{v,T}$ from v to T is $\min_{u \in V_T} |p_{v,u}|$, and u here represents the node closest to v in T . After finding $d_{v,T}$ for every v , BAERA extracts the node v_{\min} with the smallest $d_{v,T}$, i.e., $v_{\min} = \arg \min_{v \in K - V_T} d_{v,T}$ and adds $p_{v,u}$ to T .² Most importantly, to avoid constantly generating a new branch node, BAERA will choose $p_{v_{\min},u}$, i.e., let the node v_{\min} connect to u which already acted as a branch node in T , if there are multiple v_{\min} sharing the same minimal distance $d_{v_{\min},T}$. Edge Optimization Phase ends when all nodes in K are added to V_T .

Fig. 2(a) presents an example of Edge Optimization Phase, where node 1 is the root. Node 2 is first connected to node 1 with 2 edges via node d . Node 3 is then connected to d with 3 edges via nodes b and a . Node 4 is then connected to b with 2 edges via c . Afterwards, node 5 and node 6 are connected to T sequentially. For node 7 and node 8, note that $d_{7,T}$ and $d_{8,T}$ are both 4 in Fig. 1(a), and considering $p_{8,y}$ will not generate another branch node, therefore node 8 is first connected to T and then node 7 is connected to T via the created branch node s . Afterwards, node 9 and node 10 are connected to T sequentially.

Afterwards, Branch Optimization Phase re-routes the tree

²In this paper, we connect v_{\min} to T via the shortest path. Nevertheless, it is also allowed to connect v_{\min} to T with an alternate path derived according to unicast traffic engineering [4] to meet the unicast traffic requirements.

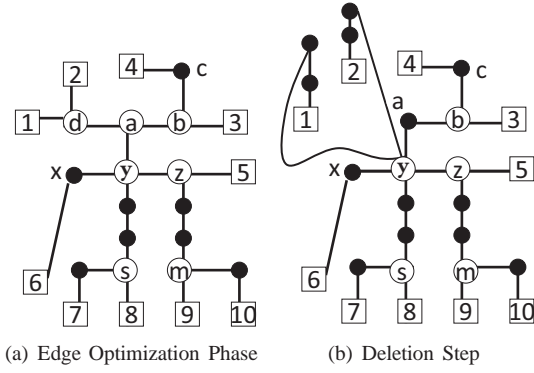


Fig. 2. An example of BAERA (refer to Fig.1)

T to reduce the number of branch nodes. Intuitively, if more branch nodes are allowed in T , the nodes in K can connect to T with shorter paths, as the plan in Edge Optimization Phase. Nevertheless, as the weight w of a branch node increases, it is necessary for a terminal node to pursue a longer path that directly connects to an existing branch node in T to avoid creating a new branch node. To address this issue, Branch Optimization Phase includes two steps: 1) Deletion Step and 2) Alternation Step. Deletion Step first tries to remove some branch nodes in T obtained from Edge Optimization Phase, and then Alternation Step tries to iteratively move each of remaining branch nodes to its neighbor node. In the above two steps, the solution T will be replaced by the new one only if its objective value $c(T) + b(T)w$ is improved (i.e., reduced).

More specifically, Deletion Step first sorts the branch nodes by the ascending order of the degree in T . In other words, a branch node owning fewer neighbor branch nodes and neighbor terminal nodes³ will be examined first because the solution has a higher chance to be improved. When a branch node v_d is removed, because T is partitioned into multiple connected components, v_d 's neighbor branch node and neighbor terminal node will correspond to different connected components. Deletion Step will re-route v to the v 's closest branch node u in another connected component via its shortest path $p_{v,u}$ to merge the two connected components.⁴ This process is repeated such that different connected components will be connected together to create a new tree. Fig. 2(b) presents an example of deleting branch node d from Fig. 2(a). After d is deleted, node 1 and node 2 are re-routed to the other connected component's node a via node y . Therefore, the number of branch nodes can be reduced when Deletion Step ends.

Afterwards, Alternation Step sorts the branch nodes in the ascending order of the degree again. This step tries to move each branch node v_a to a neighbor node v_n . For each neighbor branch node or neighbor terminal node v

³Herein, the examples for the neighbor terminal node and neighbor branch node are presented. In Fig. 2(b), node 2 is a neighbor terminal node of y because there is no other branch node or terminal node between them, while node 4 is not the neighbor terminal node of y . Node b is a neighbor branch node of y , but is not a neighbor branch node of s .

⁴If a cycle is created by adding $p_{v,u}$, the longest path between two neighbor branch nodes in the cycle can be removed.

of v_a , p_{v,v_a} is replaced by p_{v,v_n} .⁵ This step will choose the neighbor node v_n leading to the most reduction on the objective value $c(T) + b(T)w$, and each branch nodes can be moved multiple times until no neighbor node is able to reduce the objective value. The difference between Alternation Step and Deletion Step is that here every v in different connected component will connect to the same node (i.e., v_n), leading to a chance on the reduction of the edge number. Fig. 1(d) presents the result of altering branch node b to its neighbor c in Fig. 2(b). Paths $p_{y,b}$ and $p_{3,b}$ are replaced by paths $p_{y,c}$ and $p_{3,c}$ with $c(T)$ reduced by 1.

In the following, we prove that BAERA with the above two phases is a k -approximation algorithm if the optimal solution includes at least one branch node. On the other hands, when the optimal solution has no branch node, it will become a path, instead of a tree. We will discuss this case later.

Theorem 2: BAERA is a k -approximation algorithm for the BST problem.

Proof: In Edge Optimization Phase, since T is constructed by adding shortest paths to T , $c(T) = \sum_{v \in K} d_{v,T}$ as explained early in this section. Because $d_{v,T} = \min_{u \in V_T} |p_{v,u}|$ and the root node $r \in V_T$, $d_{v,T} \leq d_{v,r}$, where $d_{v,r}$ is the number of edges in the shortest path from v to r . Let T^* denote the optimal BST, and $d_{v,r}^*$ denote number of edges in the path from v to r on T^* , which may not be the shortest path between v and r in G . In other words, $d_{v,r} \leq d_{v,r}^*$. Apparently, $d_{v,r}^* \leq c(T^*)$, and thus we conclude that $c(T) = \sum_{v \in K} d_{v,T} \leq \sum_{v \in K} d_{v,r} \leq \sum_{v \in K} d_{v,r}^* \leq k * c(T^*)$ after the first phase ends. On the other hand, T cannot have more than k branch nodes because each step in this phase creates at most one branch node. Therefore, $b(T) \leq k * b(T^*)$ since $b(T^*) \geq 1$, and the tree T generated in the first phase is k -approximated. Since the second phase re-routes the tree only if the objective value $c(T) + b(T)w$ can be reduced, the tree T outputted in the second phase is also k -approximated. The theorem follows. ■

In the following, we discuss the cases when the optimal solution has no branch node, i.e., the optimal solution is a path, instead of a tree. Let P^* denote the optimal BST.

Proposition 1: If $w \leq k$, then BAERA is a $2k$ -approximation algorithm.

Proof: Denote T as the tree generated by BAERA. First, since T is constructed by adding shortest paths to T , we know that $c(T) \leq k * c(P^*)$ according to Theorem 2. Second, since BAERA includes at most one additional branch node in each iteration, there are at most $k-2$ branch nodes in T , i.e., $k-2 \leq b(T)$. In addition, since P^* connects all terminals in K , the number of edges in P^* must be at least $k-1$, i.e., $k-1 \leq c(P^*)$. Thus, we obtain $b(T)w \leq (k-2)w \leq c(P^*) * k$. Therefore, $c(T) + b(T)w \leq 2k * c(P^*)$ and BAERA is a $2k$ -approximation algorithm when $w \leq k$. The theorem follows. ■

BAERA tends to generate a solution with branch nodes. For the case with a large w , we explore another direction that leverages the Hamiltonian path to find a solution with the performance guarantee. In the following, we first introduce

⁵Any cycle created by adding p_{v,v_n} is also necessary to be removed.

the Ore's Theorem [22], and then prove that an Hamiltonian path must exist if the degree of selected nodes are large enough in Proposition 2.

Theorem 3 (Ore's Theorem): Let $G = (V, E)$ be a connected simple graph with $n \geq 3$ vertices. If for each pair of non-adjacent vertices $u, v \in V$ such that $\deg(u) + \deg(v) \geq n$, G contains a Hamiltonian cycle.

Proposition 2: Assume there exists a connected subgraph H of G and for each pair of non-adjacent vertices $u, v \in V(H)$, $\deg_H(u) + \deg_H(v) \geq |V(H)|$. If $K \subseteq V(H)$ and $|V(H)| \leq (k-1)k$, then we can find a Hamiltonian path P with $\frac{c(P)}{c(P^*)} \leq k$.

Proof: In the following, we discuss the case when $|V(H)| \geq 3$.⁶ Since H is connected and for each pair of non-adjacent vertices $u, v \in V(H)$, $\deg_H(u) + \deg_H(v) \geq |V(H)|$ holds, H has a Hamiltonian cycle C according to Ore's Theorem. Thus we can find a path P from C such that the start node and end node are in K and connects all terminal nodes which satisfies $c(P) \leq |V(H)| \leq (k-1)k$. Since $c(P^*) \geq (k-1)$ (as mentioned in Proposition 1), we obtain $\frac{c(P)}{c(P^*)} \leq \frac{(k-1)k}{k-1} = k$. The theorem follows. ■

Note that those H can be obtained by examining dense subgraphs [23] or k -cores [24]. Moreover, if such H exists, the corresponding Hamiltonian Path P could be derived by an existing algorithm [25].

Time Complexity. We first find the shortest path between any two nodes in G with Johnson's algorithm in $O(|V||E| + |V|^2 \log |V|)$ time as a pre-processing procedure for quickly lookup afterwards. The advantage is that the preprocessing only needs to be performed once but can be exploited during the construction of all BSTs afterwards. In each iteration of Edge Optimization Phase, BAERA finds $d_{v,T}$ and extracts v_{\min} in $O(k|V|)$ time, and this phase requires $O(k^2|V|)$ time to connect all terminal nodes to T .

In Branch Optimization Phase, let B denote the set of branch nodes in T . Let δ_T denote the maximal degree of a node in T , and $\delta_T \leq k$ and $\delta_T \leq \delta_G$ must hold, where δ_G is the maximal degree of a node in G . Deletion Step first sorts the branch nodes in the ascending order of the degree in T . Since $|B| \leq k-2$, the sorting requires $O(k \log k)$ time. We then build a heap for each branch node to store the shortest-path distance from other branch nodes to v in $O(k \log k)$ time. To remove a branch node v_d , it is necessary to connect each neighbor branch node and neighbor terminal node v to the existing closest branch node u in T in $O(\log k)$ time. Therefore, Deletion Step takes $O(\delta_T \log k)$ time to delete a branch node, and thus $O(k\delta_T \log k)$ for trying to delete all branch nodes. In Alternation Step, first the branch nodes are sorted in $O(k \log k)$ time. Then, BAERA tries to move each branch node in order. Note that each branch node v_a can be moved at most $O(|V|)$ times, and moving v_a to a neighbor takes $O(\delta_T)$ time. Alternation Step takes $O(k \log k + k\delta_T|V|)$ time. Therefore, the time complexity of Branch Optimization Phase is $O(k \log k + k\delta_T|V|)$, and BAERA takes $O(k^2|V| + k\delta_T|V|)$ time after the pre-processing procedure. As shown in Section V later, δ_T is usually small, and thus the time complexity of BAERA

Algorithm 1 Branch Aware Edge Reduction Algorithm (BAERA)

Require: A network $G = (V, E)$, a nonnegative value w and a terminal set K .

Ensure: A Steiner tree T .

```

1: //Edge Optimization Phase
2: Choose a terminal node  $r$  as the root
3:  $T \leftarrow \{r\}$ ,  $K \leftarrow K - \{r\}$ ,  $A(T) \leftarrow 0$ 
4: while  $K \neq \emptyset$  do
5:   for  $v \in K$  do
6:      $d_{v,T} \leftarrow$  the minimum distance from  $v$  to  $T$ 
7:      $p_{v,T} \leftarrow$  the shortest path from  $v$  to  $T$ 
8:      $S \leftarrow \{x \mid d_{x,T} = \min_{v \in K} d_{v,T}\}$ 
9:     if there exists a  $x \in S$  such that  $T \cup p_{x,T}$  does not
       generate a new branch node then
10:       $T \leftarrow T \cup p_{x,T}$ 
11:   else
12:     Choose a  $x \in S$  and  $T \leftarrow T \cup p_{x,T}$ 
13:      $K \leftarrow K - \{x\}$ 
14:  $A(T) \leftarrow c(T) + b(T)w$  //The weight of the tree  $T$ 
15:
16: //Branch Optimization Phase 1) Deletion Step
17: Obtain an order  $\sigma$  which sorts the branch nodes in the
    ascending order of the degree in  $T$ 
18: for  $v_d \in \sigma$  do
19:    $T' \leftarrow T - \{v_d\}$ 
20:   for neighbor branch node or neighbor terminal node
      $v$  of  $v_d$  do
21:     Reroute the  $v$ 's closest branch node  $u$  in another
       connected component via its shortest path  $p_{v,u}$ 
22:      $T' \leftarrow T' \cup p_{v,u}$ 
23:     if  $c(T') + b(T')w < A(T)$  then
24:        $T \leftarrow T'$  and  $A(T) \leftarrow c(T') + b(T')w$ 
25:
26: //Branch Optimization Phase 2) Alternation Step
27: Obtain an order  $\sigma$  which sorts the branch nodes in the
    ascending order of the degree in  $T$ 
28: for  $v_a \in \sigma$  do
29:    $T' \leftarrow T - \{v_a\}$ 
30:   Choose a neighbor node  $v_n$  of  $v_a$ 
31:   for neighbor branch node or neighbor terminal node
      $v$  of  $v_a$  do
32:     The shortest path  $p_{v,v_a}$  is replaced by the shortest
       path  $p_{v,v_n}$ 
33:      $T' \leftarrow T' \cup p_{v,v_n}$ 
34:     if  $c(T') + b(T')w < A(T)$  then
35:        $T \leftarrow T'$  and  $A(T) \leftarrow c(T') + b(T')w$ 
36: return  $T$  and  $A(T)$ 

```

after pre-processing is $O(k^2|V|)$. Moreover, $|V|$ in the above analysis represents an upper bound of the cost for scanning the tree T . Since the tree size is usually much smaller than $|V|$, the computation cost is actually close to $O(k^2|T|)$.

V. SIMULATION RESULTS

In this section, we evaluate BAERA in both real networks and massive synthetic networks.

⁶Note that a connected H with $|V(H)| = 2$ contains two nodes and a link between them, and the Hamiltonian path can be easily derived.

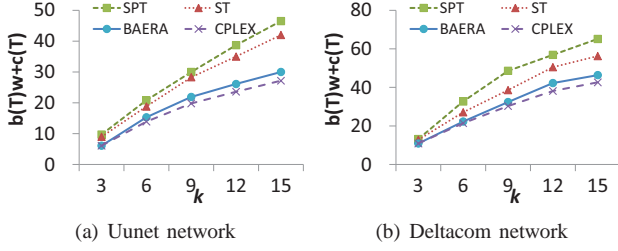


Fig. 3. Varied k in different real networks ($w = 5$)

A. Simulation Setup

The simulation is conducted in the following real networks [26]: 1) the Uunet network with 49 nodes and 84 links, and 2) the Deltacom network with 113 nodes and 183 links. Many recent SDN works [4], [16] evaluate the proposed approaches in real networks with at most hundreds of nodes. By contrast, we also evaluate our algorithm in the networks generated by Inet [27], [28] with tens of thousands of nodes to test the scalability of BAERA. In our simulation, K is chosen randomly from G .

We compare BAERA with the following algorithms: 1) the shortest-path tree algorithm (SPT), 2) a Steiner tree (ST) algorithm [8], and 3) Integer Programming solver CPLEX [29], which finds the optimal solution of the BST problem by solving the Integer Programming formulation in Section III-B. The performance metrics include: 1) the objective value of the BST problem $b(T)w + c(T)$, 2) the number of branch nodes in T , 3) the number of edges in T , and 4) the running time. All algorithms are implemented in an HP DL580 server with four Intel Xeon E7-4870 2.4 GHz CPUs and 128 GB RAM. Each simulation result is averaged over 100 samples.

B. Small Real Networks

In this subsection, we compare the performance of BAERA, ST and SPT with the optimal solutions obtained by CPLEX under different k . Since the BST problem is NP-Hard, CPLEX is able to find the optimal solutions for small instances of the BST problem, and thus we only find the optimal solutions for the Uunet and Deltacom networks. As shown in Fig. 3, the tree T grows and includes more branch nodes as k increases, because a network is inclined to generate a large tree. Nevertheless, BAERA outperforms SPT and ST in the two networks since both the edge number and the branch node number are effectively minimized. In addition, the solutions of BAERA are very close to the optimal solutions.

C. Large Synthetic Networks

In the following, we evaluate BAERA, ST and SPT in large networks with 10000 nodes generated by Inet. Fig. 4(a), Fig. 4(b), and Fig. 4(c) first discover the impact of w with k as 200. Fig. 4(a) demonstrates that the objective value $b(T)w + c(T)$ increases as w grows in all algorithms. For a larger w , BST with BAERA can effectively limit the number of the created branch nodes by slightly increasing more edges necessarily included to span all terminal nodes in K . Nevertheless, BAERA outperforms SPT and ST,

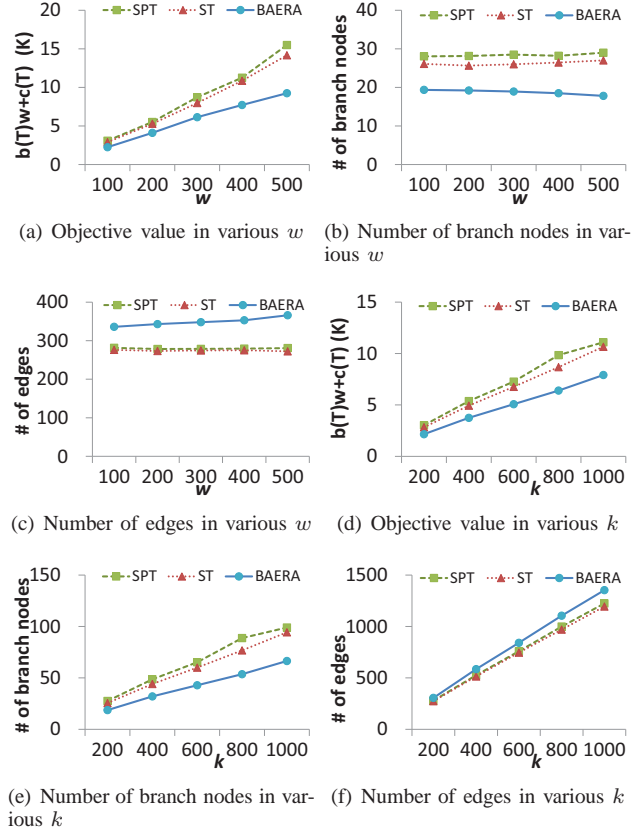


Fig. 4. Varied w and k in the synthetic network by Inet

especially for a large w , because SPT and ST focus on only the edge number and thus tend to create a tree with more branch nodes. By contrast, the number of branch nodes in the solutions obtained by BAERA is much smaller, but the edge number of BAERA is very close to ST.

Fig. 4(d), Fig. 4(e), and Fig. 4(f) evaluate the impact of k with w as 100. As shown in Fig. 4(d), the objective value $b(T)w + c(T)$ becomes larger as k increases, since more branch nodes are necessary to participate in the tree. BAERA still requires fewer branch nodes from Fig. 4(e). Moreover, the increment of the objective value in BAERA grows slower than ST and BT, showing that BAERA can further reduce the total cost in a larger k with the proposed optimization methods.

Table I and Table II evaluate the running time of BAERA with various k and different Inet graph sizes. The running time of BAERA is too small to be measured in the Uunet and Deltacom networks for arbitrary k . It demonstrates that the running time of BAERA only slightly grows for a larger k , and most instances can be solved around 6 seconds when the network has 10000 nodes. In addition, for a smaller graph, ex. 4000 nodes, BAERA takes only 1 second. Therefore, BAERA can both achieve a performance bound (i.e., k -approximation) in theory and find a good solution with small time in practice.

VI. CONCLUSIONS

Traffic engineering and flow table scalability have been studied for unicast traffic in SDN, but those issues in

TABLE I
THE RUNNING TIME OF BAERA IN DIFFERENT k ($|V|=10000$)

k	100	200	300	400
Running time (sec.)	6.064	6.418	6.816	7.430

TABLE II
THE RUNNING TIME OF BAERA IN DIFFERENT GRAPH SIZES ($k = 200$)

$ V $	4000	6000	8000	10000
Running time (sec.)	1.216	2.422	4.148	6.362

multicast SDN have not been carefully addressed. In this paper, therefore, we exploited the branch forwarding technique and proposed Branch-aware Steiner Tree (BST) for SDN. The BST problem is more difficult since it needs to jointly minimize the edge and branch node numbers in a tree, and we proved that this problem is NP-Hard and inapproximable within k . To solve this problem, we designed a k -approximation algorithm, named Branch Aware Edge Reduction Algorithm (BAERA). Simulation results manifest that the trees obtained by BAERA include fewer edges and branch nodes, compared to the shortest-path trees and Steiner trees. In addition, BAERA is efficient to be deployed in SDN because it can generate a scalable and bandwidth-efficient multicast tree in massive networks with only a few seconds.

REFERENCES

- [1] Software-defined networking (SDN) definition. [Online]. Available: <https://www.opennetworking.org/sdn-resources/sdn-definition>
- [2] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: enabling innovation in campus networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, 2008.
- [3] *OpenFlow Switch Specification*, Open Networking Foundation Std. 1.4.0, Oct. 2013.
- [4] S. Agarwal, M. Kodialam, and T. Lakshman, "Traffic engineering in software defined networks," in *IEEE Proceedings of INFOCOM*, 2013, pp. 2211–2219.
- [5] Y. Kanizo, D. Hay, and I. Keslassy, "Palette: Distributing tables in software-defined networks," in *IEEE Proceedings of INFOCOM*, 2013, pp. 545–549.
- [6] B. Fenner, M. Handley, H. Holbrook, and I. Kouvelas, "Protocol independent multicast - sparse mode (pim-sm): protocol specification (revised)," *IETF RFC 4601*, Aug. 2006.
- [7] B. Cain, S. Deering, I. Kouvelas, B. Fenner, and A. Thyagarajan, "Internet group management protocol, version 3," *IETF RFC 3376*, Oct. 2002.
- [8] H. Takahashi and A. Matsuyama, "An approximate solution for the Steiner problem in graphs," *Mathematica Japonicae*, vol. 24, pp. 571–577, 1980.
- [9] D.-N. Yang and W. Liao, "Protocol design for scalable and adaptive multicast for group communications," in *IEEE International Conference on Network Protocols*, 2008, pp. 33–42.
- [10] —, "Optimal state allocation for multicast communications with explicit multicast forwarding," *IEEE Transactions on Parallel and Distributed Systems*, vol. 19, no. 4, pp. 476–488, 2008.
- [11] J. Tian and G. Neufeld, "Forwarding state reduction for sparse mode multicast communication," in *IEEE Proceedings of INFOCOM*, 1998, pp. 711–719.
- [12] I. Stoica, T. Ng, and H. Zhang, "Reunite: A recursive unicast approach to multicast," in *IEEE Proceedings of INFOCOM*, 2000, pp. 1644–1653.
- [13] T. Wong and R. Katz, "An analysis of multicast forwarding state scalability," in *IEEE Proceedings of International Conference on Network Protocols*, 2000, pp. 105–115.
- [14] G. Robins and A. Zelikovsky, "Improved Steiner tree approximation in graphs," in *Proceedings of the eleventh annual ACM-SIAM symposium on Discrete Algorithms*, 2000, pp. 770–779.
- [15] J. Sushant, K. Alok, M. Subhasree, O. Joon, P. Leon, S. Arjun, V. Subbaiah, W. Jim, Z. Junlan, Z. Min, Z. Jon, H. Urs, S. Stephen, and V. Amin, "B4: experience with a globally-deployed software defined wan," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, pp. 3–14, 2013.
- [16] Z. A. Qazi, C. C. Tu, L. Chiang, R. Miao, V. Sekar, and M. Yu, "Simple-fying middlebox policy enforcement using sdn," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, pp. 27–38, 2013.
- [17] J. Mueller, A. Wierz, and T. Magedanz, "Scalable on-demand network management module for software defined telecommunication networks," in *IEEE Proceedings of SDN for Future Networks and Services*, Nov. 2013, pp. 1–6.
- [18] B.-S. Lee, R. Kanagavelu, and K. M. M. Aung, "An efficient flow cache algorithm with improved fairness in software-defined data center networks," in *IEEE 2nd International Conference on Cloud Networking*, 2013, pp. 18–24.
- [19] D.-N. Yang and W. Liao, "On bandwidth-efficient overlay multicast," *IEEE Transactions on Parallel and Distributed Systems*, vol. 18, no. 11, pp. 1503–1515, 2007.
- [20] E. Aharoni and R. Cohen, "Restricted dynamic Steiner trees for scalable multicast in datagram networks," *IEEE/ACM Transactions on Networking*, vol. 6, no. 3, pp. 286–297, 1998.
- [21] L.-H. Huang, H.-J. Hung, C.-C. Lin, and D.-N. Yang, "Scalable Steiner tree for multicast communications in software-defined networking," *CoRR*, vol. abs/1404.3219, 2014.
- [22] O. Ore, "Note on hamilton circuits," *American Mathematical Monthly*, vol. 67, no. 1, p. 55, 1960.
- [23] S. Khuller and B. Saha, "On finding dense subgraphs," in *ICALP*, 2009, pp. 597–608.
- [24] S. B. Seidman, "Network structure and minimum degree," *Social Networks*, vol. 5, no. 3, pp. 269–287, 1983.
- [25] E. Palmer, "The hidden algorithm of ore's theorem on hamiltonian cycles," *Computers & Mathematics with Applications*, vol. 34, no. 11, pp. 113 – 119, 1997.
- [26] The internet topology zoo. [Online]. Available: <http://www.topology-zoo.org/dataset.html>
- [27] H. Tangmunarunkit, R. Govindan, S. Jamin, S. Shenker, and W. Willinger, "Network topology generators: degree-based vs. structural," *ACM SIGCOMM Computer Communication Review*, vol. 32, no. 4, pp. 147–159, 2002.
- [28] Inet topology generator. [Online]. Available: <http://topology.eecs.umich.edu/inet/>
- [29] IBM ILOG CPLEX. [Online]. Available: <http://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/>

Scalable Steiner Tree for Multicast Communications in Software-Defined Networking

Liang-Hao Huang, Hui-Ju Hung, Chih-Chung Lin, and De-Nian Yang

Academia Sinica, Taipei, Taiwan

{lhhuang, hjhung, chchlin, dnyang}@iis.sinica.edu.tw

Abstract—Software-Defined Networking (SDN) enables flexible network resource allocations for traffic engineering, but at the same time the scalability problem becomes more serious since traffic is more difficult to be aggregated. Those crucial issues in SDN have been studied for unicast but have not been explored for multicast traffic, and addressing those issues for multicast is more challenging since the identities and the number of members in a multicast group can be arbitrary. In this paper, therefore, we propose a new multicast tree for SDN, named Branch-aware Steiner Tree (BST). The BST problem is difficult since it needs to jointly minimize the numbers of the edges and the branch nodes in a tree, and we prove that it is NP-Hard and inapproximable within k , which denotes the number of group members. We further design an approximation algorithm, called Branch Aware Edge Reduction Algorithm (BAERA), to solve the problem. Simulation results demonstrate that the trees obtained by BAERA are more bandwidth-efficient and scalable than the shortest-path trees and traditional Steiner trees. Most importantly, BAERA is computation-efficient to be deployed in SDN since it can generate a tree on massive networks in small time.

Index Terms—SDN, multicast, NP-Hard, traffic engineering, scalability

I. INTRODUCTION

Software-Defined Networking (SDN) is an emerging architecture that is manageable, dynamic, cost-effective, and adaptable, making it ideal for the high-bandwidth, huge data, and dynamic nature of numerous network services [?]. This novel architecture decouples the network control and forwarding functions. It enables the network control to become directly programmable and the underlying infrastructure to be abstracted for varied applications. The OpenFlow protocol has been recognized as a crucial element for building SDN solutions [?], [?], [?].

SDN comprises two main components: SDN controller (SDN-C) and SDN forwarding element (SDN-FE) [?]. Compared with the traditional shortest-path routing, SDN-C enables the centralized computation on unicast routing for traffic engineering [?] to improve the network throughput. Nevertheless, since the routing paths no longer need to be the shortest ones, the paths can be distributed flexibly inside the network and thus are more difficult to be aggregated in the flow table of SDN-FE, and the scalability has been regarded as a serious issue to deploy SDN in a large network [?], [?].

Multicast is an efficient technique for point-to-multipoint (P2M) and multipoint-to-multipoint (M2M) communications because it exploits a tree, instead of disjoint paths, in the routing of the traffic. Current multicast standard on Internet, i.e., PIM-SM [?], employs a shortest-path tree to connect the

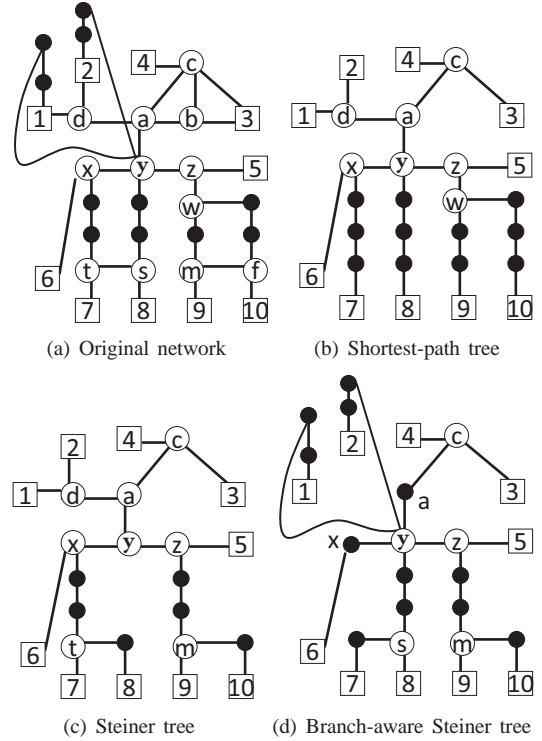


Fig. 1. An example of multicast tree

terminal nodes in a multicast group, where a terminal node is a designated router connecting to a LAN with at least one user client joining the group [?]. Traffic engineering is difficult to be supported in a shortest-path tree since the path from the root, i.e., the traffic source in P2M or the rendezvous point in M2M in PIM-SM, to each destination in the tree is still the shortest path. By contrast, a Steiner Tree (ST) [?] in Graph Theory is more promising because it minimizes the network resource consumption, i.e., the number of edges in a tree, required for a multicast group. However, finding an ST is more computation intensive and thus is difficult to be deployed as a distributed protocol on Internet. By contrast, now it becomes feasible by first finding an ST in SDN-C and then storing the forwarding information in the group tables of SDN-FEs on the tree.

Similar to unicast traffic engineering in SDN, multicast traffic engineering also suffers from the scalability problem since each SDN-FE in the tree needs to store a forwarding entry in the group table for each multicast group. Nevertheless, the scalability problem for multicast communications

is even more serious since the number of possible multicast group is $O(2^n)$, where n is the number of nodes in a network, and the number of possible unicast connections is $O(n^2)$. To remedy this issue, a promising way is to exploit the *branch forwarding technique* [?], [?], [?], [?], [?], which stores the entries in only the branch nodes, instead of every node, of a multicast tree. More specifically, a branch node in a tree is the node with at least three incident edges, such as white circle nodes in Fig. 1, and the square nodes are the terminal nodes. To minimize the total number of edges in an ST, the path connecting two neighboring branch nodes (such as nodes c and y in Fig. 1(d)) needs to be the shortest path between them. Note that an ST is not a shortest-path tree because the branch nodes can be located anywhere in the network. This branch forwarding technique can remedy the multicast scalability problem since packets are forwarded in a unicast tunnel from the logic port of a branch node in SDN-FE [?]. In other words, all nodes in the path (such as black circle nodes in Fig. 1) exploit unicast forwarding in the tunnel and are no longer necessary to maintain a forwarding entry for the multicast group.

To effectively address the multicast scalability problem in SDN, it is crucial to minimize the number of branch nodes in a tree. However, this important factor has not been considered in ST. In this paper, therefore, we propose a new multicast tree for SDN, named *Branch-aware Steiner Tree (BST)*. The objective of BST problem is to minimize the summation of the number of edges and the number of branch nodes in the tree, where a branch node can be assigned a higher weight to further improve the scalability. Fig. 1 presents an illustrative example with the weight of each branch node set as 20. Square nodes are the terminal nodes that are required to be connected in a tree, while the black and white circle nodes are the other nodes in the network. Fig. 1(a) is the network topology. The shortest-path tree in Fig. 1(b) includes 27 edges and 7 branch nodes with the total cost of the tree as $27 + 7 \times 20 = 167$. The Steiner tree in Fig. 1(c) has 23 edges and 8 branch nodes with the total cost as $23 + 8 \times 20 = 183$. By contrast, Fig. 1(d) presents the BST with 26 edges and 5 branch nodes and the total cost as $26 + 5 \times 20 = 126$. Therefore, compared with the shortest-path trees on Internet, BST effectively reduces the network resource consumption by minimizing the number of edges in the tree. Compared with ST, more BSTs can be supported in SDN since the number of branch nodes is effectively minimized.

Finding an BST is very challenging. The ST problem is NP-Hard but can be approximated within ratio 1.55 [?] and is thus in APX of complexity theory. In other words, there exists an approximation algorithm for ST that can find a tree with the total cost at most 1.55 times of the optimal solution. By contrast, we prove that BST is NP-Hard but cannot be approximated within k , which denotes the number of terminal nodes in a multicast group. In other words, the BST problem is more difficult to be approximated. To effectively solve BST, we propose a k -approximation algorithm, named *Branch Aware Edge Reduction Algorithm (BAERA)*, that can be deployed in SDN-C. BAERA includes two phases, Edge Optimization Phase and Branch Optimization Phase,

to effectively minimize the number of edges and branch nodes. Since no $(k^{1-\epsilon})$ -approximation algorithm exists in BST for arbitrarily small $\epsilon > 0$, BAERA achieves the best approximation ratio.

The rest of this paper is organized as follows. Section II briefly summarizes the literature on SDN traffic engineering, SDN flow table scalability, multicast scalability, and the Steiner tree. Section III formally presents the problem formulation with Integer Programming and the hardness result. We design a k -approximation algorithm in Section IV, and Section V presents the simulation results to evaluate the performance of the proposed algorithm in real networks. We conclude this paper in Section VI.

II. RELATED WORKS

Previous works have extensively explored the issues on traffic engineering and flow table scalability for *unicast traffic* in SDN. Mckeown et al. [?] pointed out that OpenFlow can be deployed with heterogeneous switches. Sushant et al. [?] shared their experience of SDN development for the private WAN of Google Inc. Qazi et al. [?] proposed a new system design using SDN for the middleboxes (e.g., firewalls, VPN gateways, proxies). Agarwal et al. [?] considered the incremental deployment of traffic engineering in the case where a SDN-C controls only a few SDN-FEs in the network, and the rest of the network adopts a standard routing protocol, such as OSPF. The merits of traffic engineering brought by only a limited number of SDN-capable nodes are demonstrated. Mueller et al. [?] presented a cross-layer framework in SDN, which integrates a novel dynamic traffic engineering approach with an adaptive network management, to bridge the gap between the network and application layers for overall system optimizations.

On the other hand, flow table scalability is crucial to enable a large-scale deployment of SDN. For unicast traffic, Kanizo et al. [?] pointed out that the restriction on table sizes is the major bottleneck in SDN and proposed a framework, called Palette, to decompose a large SDN table into small ones and then distribute them across the network. Lee et al. [?] observed that Data Center traffic frequently meets few elephant flows and a lot of mice flows. However, elephant flows are inclined to be evicted because of the limited flow table sizes. They proposed a differential flow cache framework that uses a hash-based cache placement and localized Least Recently Used (LRU)-based replacement to reduce the loss of elephant flows.

The scalability issue is more serious in multicast, and the previous works [?], [?], [?], [?], [?] have demonstrated that the branch forwarding technique is a promising way since forwarding from a branch node to a neighbor branch node or terminal node can exploit the existing unicast tunneling technique, and tunneling can be facilitated in SDN with logic ports specified in the group table [?]. In other words, the intermediate nodes between two neighbor branch routers no longer need to store a multicast forwarding entry for the tree. However, the above works were designed for shortest-path trees and did not explore the possibility of more flexible multicast routing. On the other hand, Steiner tree [?] can effectively minimize the bandwidth consumption in a network, but so far it is not adopted on Internet since finding

the optimal Steiner tree is more computation intensive and thus difficult to be deployed as a distributed protocol. To remedy this issue, overlay Steiner trees [?], [?] for P2P environments are proposed, where only the terminal nodes can act as branch nodes. Nevertheless, the merit of traffic engineering from the above work is limited since no other router can act as the branch node to reduce the bandwidth consumption. Moreover, multicast scalability is not studied in the above works. Therefore, the above works are difficult for bandwidth-efficient and scalable multicast in SDN.

III. PRELIMINARIES

A. Problem Formulation

In this paper, we propose a scalable and bandwidth-efficient multicast tree for SDN, called *Branch-aware Steiner Tree (BST)*. This paper aims to minimize the bandwidth consumption (i.e., the total number of links/edges) and the number of forwarding entries maintained for the multicast group (i.e., the total branch nodes). Therefore, the BST problem is to find a tree connecting a given set of terminal nodes such that the sum of the number of edges and the number of branch nodes is minimized, where a branch node can be assigned a larger weight w to ensure a higher scalability.¹

Definition 1: Consider a network $G(V, E)$, where V and E denote the set of nodes and edges, respectively. Given $G(V, E)$, a terminal node set $K \subseteq V$, and a non-negative value w , the BST problem is to find a tree T spanning the terminal node set K such that $c(T) + b(T)w$ is minimized, where $c(T)$ is the number of edges on T , and $b(T)$ is the number of branch nodes (i.e., nodes with the degree at least 3 on T).

In BST, a network operator can increase the scalability of multicast in SDN by assigning a larger weight w for branch nodes. Compared with ST, $c(T)$ may slightly increase, but much fewer branch nodes will be selected in T . Compared with the shortest-path trees adopted on Internet currently, BST allows more flexible routing of a tree and thus can effectively reduce the network resource consumption and improve the scalability in SDN.

In the following, we first formulate the BST problem as an Integer Programming problem. Afterwards, we show that the BST problem is very challenging in complexity theory by proving that it is NP-Hard and not able to be approximated within k^c for every $c < 1$.

B. Integer Programming

Let N_v denote the set of neighbor nodes of v in G , and u is in N_v if $e_{u,v}$ is an edge from u to v in E . Let any terminal node r act as the root of T , i.e., the source, and the destination set L contains the other terminals in K , i.e., $L = K - \{r\}$. The output tree T needs to ensure that there is only one path in T from r to every node in L . To achieve this goal, our problem includes the following binary decision variables. Let binary variable $\pi_{l,u,v}$ denote if edge $e_{u,v}$ is

in the path from r to a destination node l in L . Let binary variable $\varepsilon_{u,v}$ denote if edge $e_{u,v}$ is in T , where $\varepsilon_{u,v} = \varepsilon_{v,u}$. Let binary variable β_v denote if v is a branch node in T . Intuitively, when we are able to find the path from r to each destination node l with $\pi_{l,u,v} = 1$ on every edge $e_{u,v}$ in the path, the routing of the tree with $\varepsilon_{u,v} = 1$ for every edge $e_{u,v}$ in T can be constructed with the union of the paths from r to all destination nodes in L , and every branch node v in T with $\beta_v = 1$ in T can be identified accordingly.

Most importantly, to guarantee that the union of the paths is a tree, i.e., a subgraph without any cycle, the objective function of our Integer Programming formulation (IP) is as follows.

$$\min \sum_{e_{u,v} \in E} \varepsilon_{u,v} + \sum_{v \in V} w \times \beta_v.$$

If the tree T contains any cycle, T is not optimal since we are able to remove at least one edge from the cycle to reduce the objective value, and ensure that there still exist a path from r to every destination node l in L . To find $\varepsilon_{u,v}$ and β_v from $\pi_{l,u,v}$, our IP formulation includes the following constraints.

$$\sum_{v \in N_r} \pi_{l,r,v} - \sum_{v \in N_r} \pi_{l,v,r} = 1, \forall l \in L, \quad (1)$$

$$\sum_{u \in N_l} \pi_{l,u,l} - \sum_{u \in N_l} \pi_{l,l,u} = 1, \forall l \in L, \quad (2)$$

$$\sum_{v \in N_u} \pi_{l,v,u} = \sum_{v \in N_u} \pi_{l,u,v}, \quad (3)$$

$$\forall l \in L, \forall u \in V, u \neq l, u \neq r, \quad (4)$$

$$\pi_{l,u,v} \leq \varepsilon_{u,v}, \forall l \in L, \forall e_{u,v} \in E, \quad (4)$$

$$\frac{1}{|N_u|} \left(-2 + \sum_{v \in N_u} \varepsilon_{u,v} \right) \leq \beta_u, \forall u \in V. \quad (5)$$

The first three constraints, i.e., (1), (2), and (3), are the flow-continuity constraints to find the path from r to every destination node l in L . More specifically, r is the flow source, i.e., the source of the path to every destination node l , and constraint (1) states that the net outgoing flow from r is one, implying that at least one edge $e_{r,v}$ from r to any neighbor node v needs to be selected with $\pi_{l,r,v} = 1$. Note that here decision variables $\pi_{l,r,v}$ and $\pi_{l,v,r}$ are two different variables because the flow is directed. On the other hand, every destination node l is the flow destination, and constraint (2) ensures that the net incoming flow to l is one, implying that at least one edge $e_{u,l}$ from any neighbor node u to l must be selected with $\pi_{l,u,l} = 1$. For every other node u , constraint (3) guarantees that u is either located in the path or not. If u is located in the path, both the incoming flow and outgoing flow for u are at least one, indicating that at least one binary variable $\pi_{l,v,u}$ is 1 for the incoming flow, and at least one binary variable $\pi_{l,u,v}$ is 1 for the outgoing flow. Otherwise, both $\pi_{l,v,u}$ and $\pi_{l,u,v}$ are 0. Note that the objective function will ensure that $\pi_{l,v,u} = 1$ for at most one neighbor node v to achieve the minimum cost. In other words, both the incoming flow and outgoing flow among u and v cannot exceed 1.

Constraints (4) and (5) are formulated to find the routing of the tree and its corresponding branch nodes, i.e., $\varepsilon_{u,v}$ and β_v . Constraint (4) states that $\varepsilon_{u,v}$ must be 1 if edge $e_{u,v}$ is included in the path from r to at least one l , i.e.,

¹Note that this problem can be simply extended to support different weights on each edges and each nodes. For example, a congested edge or a node with the group table almost full can be assigned a higher weight.

$\pi_{l,u,v} = 1$. The tree T is the union of the paths from r to all destination nodes. Note that here $\varepsilon_{u,v}$ and $\varepsilon_{v,u}$ represent the same binary decision variable because T is not directed. In other words, $\varepsilon_{u,v} = 1$ if edge $e_{u,v}$ is in a path (i.e., a directed flow) from either direction. The last constraint is the most crucial one. For each node u , if the degree of u is at least 3 in T , $\sum_{v \in N_u} \varepsilon_{u,v} \geq 3$ holds, and thus the left-hand-side of constraint (5) becomes positive, thereby enforcing that $\beta_u = 1$ and u acts as a branch node. Otherwise, the left-hand-side of constraint (5) is 0 or negative, allowing β_u to be 0 to minimize the cost in the objective function. In this case, node u is not a branch node in T .

C. Hardness result

The BST problem is NP-Hard because it is equivalent to the ST problem when w is 0. In other words, the ST problem is a special case of the BST problem. However, the BST is much more challenging because the ST problem can be approximated within ratio 1.55 [?] and is thus in APX in complexity theory, but we find out that BST is much more difficult to be approximated. The following theorem proves that the BST problem cannot be approximated within k^c for every $c < 1$, by a gap-introducing reduction from the Hamiltonian path problem, which determines whether there exists a path going through every node on a graph exactly once.

Theorem 1: For any $\epsilon > 0$, there exists no $k^{1-\epsilon}$ approximation algorithm for the BST problem, assuming $P \neq NP$.

Proof: We prove the theorem with the gap-introducing reduction from the Hamiltonian path problem. For an instance $G_H(V_H, E_H)$ of the Hamiltonian path problem with any node v on G_H , we build an instance of the BST problem on $G(V, E)$, such that

- if a Hamiltonian path starting at v exists in G_H , $\text{OPT}(G) \leq 2h$, and
- if no Hamiltonian path starting at v exists in G_H , $\text{OPT}(G) > 2hk^{1-\epsilon}$

, where h is the number of nodes in G and $\text{OPT}(G)$ is the optimal solution of G for the BST problem.

We first detail how to build the instance of the BST problem from the Hamiltonian path problem. For any given G_H , we construct a new graph G which consists of n^p copies of G_H , where n is the number of nodes in G_H and p is the smallest integer following $p \geq \frac{2}{\epsilon}$. One additional node x is added to G to connect to the node v of each of the n^p copies. The K is set to $V - \{x\}$ and w is set to h , where h is the number of nodes in G , i.e., $h = (n^p) \times n + 1$.

If G_H has a Hamiltonian path starting at v , consider a tree rooted at x , which includes 1) the edges between x and v of all copies and 2) the edges on the Hamiltonian path of all copies. The tree is a feasible solution of the BST problem with only one branch node x , and it can act as an upper bound of the BST in G . Thus, $\text{OPT}(G) \leq h + (h - 1) < 2h$. On the other hand, if G_H does not have a Hamiltonian path starting at v , there must exist at least one additional branch node in each copy of G . Hence, $\text{OPT}(G) > hn^p \geq 2hn^{p-1} = 2h(n^{p+1})^{\frac{p-1}{p+1}} \geq 2h(n^{p+1})^{1-\epsilon} = 2hk^{1-\epsilon}$. Since ϵ can be arbitrarily small, for any $\epsilon > 0$, there is no $k^{1-\epsilon}$ approximation algorithm for the BST problem, assuming $P \neq NP$. The theorem follows. ■

IV. ALGORITHM DESIGN

For BST, the shortest-path tree is not a good solution since the shortest path for each node v in K is constructed individually. With the aim to minimize the number of the edges, substituting the shortest path of v with a longer path can reduce the total edge number when the path mostly overlaps with the path to another node $v' \in K$ [?]. Therefore, it is expected that aggregating two paths that share more common edges can effectively reduce the number of edges in T . Nevertheless, aggregating two paths that partially overlap will generate a new branch node, and more branch nodes are inclined to be created when more paths are aggregated. Without considering the number of branch nodes created, the solution quality may deteriorate even though the number of edges in T is effectively reduced. In the following, therefore, we propose a k -approximation algorithm for BST, called *Branch Aware Edge Reduction Algorithm* (BAERA), to jointly minimize the numbers of edges and branch nodes in T . As Theorem 1 proves that no $(k^{1-\epsilon})$ -approximation algorithm for any $\epsilon > 0$ for the BST problem, BAERA achieves the best approximation ratio. Due to space constraint, the pseudo code is presented in [?].

BAERA includes two phases: 1) Edge Optimization Phase and 2) Branch Optimization Phase. In the first phase, BAERA iteratively chooses and adds a terminal node in K to the solution tree $T(V_T, E_T)$ for constructing a basic BST, where V_T and E_T denote the nodes and edges currently in T , respectively at each iteration. Initially, a random root node is added to V_T . Afterwards, for each terminal node $v \in K$ that is not in V_T , BAERA first finds the minimal distance $d_{v,T}$ from v to T . Precisely, let $p_{v,u}$ denote the shortest path from v to u on the network G , and $|p_{v,u}|$ is the number of edges in $p_{v,u}$. The minimal distance $d_{v,T}$ from v to T is $\min_{u \in V_T} |p_{v,u}|$, and u here represents the node closest to v in T . After finding $d_{v,T}$ for every v , BAERA extracts the node v_{\min} with the smallest $d_{v,T}$, i.e., $v_{\min} = \arg \min_{v \in K - V_T} d_{v,T}$ and adds $p_{v,u}$ to T .² Most importantly, to avoid constantly generating a new branch node, BAERA will choose $p_{v_{\min},u}$, i.e., let the node v_{\min} connect to u which already acted as a branch node in T , if there are multiple v_{\min} sharing the same minimal distance $d_{v_{\min},T}$. Edge Optimization Phase ends when all nodes in K are added to V_T .

Fig. 2(a) presents an example of Edge Optimization Phase, where node 1 is the root. Node 2 is first connected to node 1 with 2 edges via node d . Node 3 is then connected to d with 3 edges via nodes b and a . Node 4 is then connected to b with 2 edges via c . Afterwards, node 5 and node 6 are connected to T sequentially. For node 7 and node 8, note that $d_{7,T}$ and $d_{8,T}$ are both 4 in Fig. 1(a), and considering $p_{8,y}$ will not generate another branch node, therefore node 8 is first connected to T and then node 7 is connected to T via the created branch node s . Afterwards, node 9 and node 10 are connected to T sequentially.

Afterwards, Branch Optimization Phase re-routes the tree

²In this paper, we connect v_{\min} to T via the shortest path. Nevertheless, it is also allowed to connect v_{\min} to T with an alternate path derived according to unicast traffic engineering [?] to meet the unicast traffic requirements.

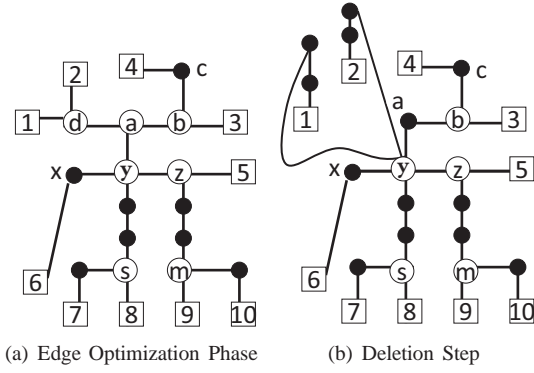


Fig. 2. An example of BAERA (refer to Fig.1)

T to reduce the number of branch nodes. Intuitively, if more branch nodes are allowed in T , the nodes in K can connect to T with shorter paths, as the plan in Edge Optimization Phase. Nevertheless, as the weight w of a branch node increases, it is necessary for a terminal node to pursue a longer path that directly connects to an existing branch node in T to avoid creating a new branch node. To address this issue, Branch Optimization Phase includes two steps: 1) Deletion Step and 2) Alternation Step. Deletion Step first tries to remove some branch nodes in T obtained from Edge Optimization Phase, and then Alternation Step tries to iteratively move each of remaining branch nodes to its neighbor node. In the above two steps, the solution T will be replaced by the new one only if its objective value $c(T) + b(T)w$ is improved (i.e., reduced).

More specifically, Deletion Step first sorts the branch nodes by the ascending order of the degree in T . In other words, a branch node owning fewer neighbor branch nodes and neighbor terminal nodes³ will be examined first because the solution has a higher chance to be improved. When a branch node v_d is removed, because T is partitioned into multiple connected components, v_d 's neighbor branch node and neighbor terminal node will correspond to different connected components. Deletion Step will re-route v to the v 's closest branch node u in another connected component via its shortest path $p_{v,u}$ to merge the two connected components.⁴ This process is repeated such that different connected components will be connected together to create a new tree. Fig. 2(b) presents an example of deleting branch node d from Fig. 2(a). After d is deleted, node 1 and node 2 are re-routed to the other connected component's node a via node y . Therefore, the number of branch nodes can be reduced when Deletion Step ends.

Afterwards, Alternation Step sorts the branch nodes in the ascending order of the degree again. This step tries to move each branch node v_a to a neighbor node v_n . For each neighbor branch node or neighbor terminal node v

³Herein, the examples for the neighbor terminal node and neighbor branch node are presented. In Fig. 2(b), node 2 is a neighbor terminal node of y because there is no other branch node or terminal node between them, while node 4 is not the neighbor terminal node of y . Node b is a neighbor branch node of y , but is not a neighbor branch node of s .

⁴If a cycle is created by adding $p_{v,u}$, the longest path between two neighbor branch nodes in the cycle can be removed.

of v_a , p_{v,v_a} is replaced by p_{v,v_n} .⁵ This step will choose the neighbor node v_n leading to the most reduction on the objective value $c(T) + b(T)w$, and each branch nodes can be moved multiple times until no neighbor node is able to reduce the objective value. The difference between Alternation Step and Deletion Step is that here every v in different connected component will connect to the same node (i.e., v_n), leading to a chance on the reduction of the edge number. Fig. 1(d) presents the result of altering branch node b to its neighbor c in Fig. 2(b). Paths $p_{y,b}$ and $p_{3,b}$ are replaced by paths $p_{y,c}$ and $p_{3,c}$ with $c(T)$ reduced by 1.

In the following, we prove that BAERA with the above two phases is a k -approximation algorithm if the optimal solution includes at least one branch node. On the other hands, when the optimal solution has no branch node, it will become a path, instead of a tree. We will discuss this case later.

Theorem 2: BAERA is a k -approximation algorithm for the BST problem.

Proof: In Edge Optimization Phase, since T is constructed by adding shortest paths to T , $c(T) = \sum_{v \in K} d_{v,T}$ as explained early in this section. Because $d_{v,T} = \min_{u \in V_T} |p_{v,u}|$ and the root node $r \in V_T$, $d_{v,T} \leq d_{v,r}$, where $d_{v,r}$ is the number of edges in the shortest path from v to r . Let T^* denote the optimal BST, and $d_{v,r}^*$ denote number of edges in the path from v to r on T^* , which may not be the shortest path between v and r in G . In other words, $d_{v,r} \leq d_{v,r}^*$. Apparently, $d_{v,r}^* \leq c(T^*)$, and thus we conclude that $c(T) = \sum_{v \in K} d_{v,T} \leq \sum_{v \in K} d_{v,r} \leq \sum_{v \in K} d_{v,r}^* \leq k * c(T^*)$ after the first phase ends. On the other hand, T cannot have more than k branch nodes because each step in this phase creates at most one branch node. Therefore, $b(T) \leq k * b(T^*)$ since $b(T^*) \geq 1$, and the tree T generated in the first phase is k -approximated. Since the second phase re-routes the tree only if the objective value $c(T) + b(T)w$ can be reduced, the tree T outputted in the second phase is also k -approximated. The theorem follows. ■

In the following, we discuss the cases when the optimal solution has no branch node, i.e., the optimal solution is a path, instead of a tree. Let P^* denote the optimal BST.

Proposition 1: If $w \leq k$, then BAERA is a $2k$ -approximation algorithm.

Proof: Denote T as the tree generated by BAERA. Since BAERA includes at most one additional branch nodes in each iteration, there are at most $k - 2$ branch nodes in T , i.e., $b(T) \leq k - 2$. In addition, since T is constructed by adding shortest paths to T , we know that $c(T) \leq k * c(P^*)$ as in Theorem ??.

Since $w \leq k$ and $b(T) \leq k - 2$, thus $c(T) + b(T)w \leq c(T) + k * (k - 2)$. According to Theorem ??, we know $c(T) \leq k * c(P^*)$. Clearly, $k - 2 \leq c(P^*)$. Therefore $c(T) + b(T)w \leq k * c(P^*) + k * c(P^*) \leq 2k * c(P^*)$. Thus BAERA is a $2k$ -approximation algorithm. ■

Theorem 3 (Ore's Theorem): Let $G = (V, E)$ be a connected simple graph with $n \geq 3$ vertices. If G has the property that for each pair of non-adjacent vertices $u, v \in V$,

⁵Any cycle created by adding p_{v,v_n} is also necessary to be removed.

we have that $\deg(u) + \deg(v) \geq n$ then G contains a Hamiltonian cycle.

Proposition 2: Let $S \subseteq V - K$ be an independent set of G . Suppose the subgraph H induced by $S \cup K$ is connected and for each pair of non-adjacent vertices $u, v \in V(H)$, $\deg_H(u) + \deg_H(v) \geq |V(H)|$ then we can find a feasible path P with $c(P)$ at most $2k - 2$.

Proof: We may assume $V(H) \geq 3$. Since H is connected and for each pair of non-adjacent vertices $u, v \in V(H)$, $\deg_H(u) + \deg_H(v) \geq |V(H)|$. By Ore's Theorem, H has a Hamiltonian cycle C . Thus we can find a path P in C that start node and end node are in K and connects all terminal nodes. Since S is an independent set, thus for each terminal node u and its neighbor terminal node v , the distance of u and v in P is at most 2. Hence the weight $c(P)$ of path P at most $2k - 2$.

If P is not a feasible path, then there exists terminal node x and its neighbor terminal node y does not connect by the shortest path. So the path connects x and y of length 2 and x and y are adjacent in G . Then we can replace this path by edge xy and reduce the weight $c(T)$. Repeat this process until P is a feasible path. Thus we can find a feasible path P with $c(P)$ at most $2k - 2$. ■

Time Complexity. We first find the shortest path between any two nodes in G with Johnson's algorithm in $O(|V||E| + |V|^2 \log |V|)$ time as a pre-processing procedure for quickly lookup afterwards. The advantage is that the preprocessing only needs to be performed once but can be exploited during the construction of all BSTs afterwards. In each iteration of Edge Optimization Phase, BAERA finds $d_{v,T}$ and extracts v_{\min} in $O(k|V|)$ time, and this phase requires $O(k^2|V|)$ time to connect all terminal nodes to T .

In Branch Optimization Phase, let B denote the set of branch nodes in T . Let δ_T denote the maximal degree of a node in T , and $\delta_T \leq k$ and $\delta_T \leq \delta_G$ must hold, where δ_G is the maximal degree of a node in G . Deletion Step first sorts the branch nodes in the ascending order of the degree in T . Since $|B| \leq k - 2$, the sorting requires $O(k \log k)$ time. We then build a heap for each branch node to store the shortest-path distance from other branch nodes to v in $O(k \log k)$ time. To remove a branch node v_d , it is necessary to connect each neighbor branch node and neighbor terminal node v to the existing closest branch node u in T in $O(\log k)$ time. Therefore, Deletion Step takes $O(\delta_T \log k)$ time to delete a branch node, and thus $O(k \delta_T \log k)$ for trying to delete all branch nodes. In Alternation Step, first the branch nodes are sorted in $O(k \log k)$ time. Then, BAERA tries to move each branch node in order. Note that each branch node v_a can be moved at most $O(|V|)$ times, and moving v_a to a neighbor takes $O(\delta_T)$ time. Alternation Step takes $O(k \log k + k \delta_T |V|)$ time. Therefore, the time complexity of Branch Optimization Phase is $O(k \log k + k \delta_T |V|)$, and BAERA takes $O(k^2 |V| + k \delta_T |V|)$ time after the pre-processing procedure. As shown in Section V later, δ_T is usually small, and thus the time complexity of BAERA after pre-processing is $O(k^2 |V|)$. Moreover, $|V|$ in the above analysis represents an upper bound of the cost for scanning the tree T . Since the tree size is usually much smaller than $|V|$, the computation cost is actually close to $O(k^2 |T|)$.

Algorithm 1 Branch Aware Edge Reduction Algorithm (BAERA)

Require: A network $G = (V, E)$, a nonnegative value w and a terminal set K .

Ensure: A Steiner tree T .

```

1: //Edge Optimization Phase
2: Choose a terminal node  $r$  as the root
3:  $T \leftarrow \{r\}$ ,  $K \leftarrow K - \{r\}$ ,  $A(T) \leftarrow 0$ 
4: while  $K \neq \emptyset$  do
5:   for  $v \in K$  do
6:      $d_{v,T} \leftarrow$  the minimum distance from  $v$  to  $T$ 
7:      $p_{v,T} \leftarrow$  the shortest path from  $v$  to  $T$ 
8:      $S \leftarrow \{x \mid d_{x,T} = \min_{v \in K} d_{v,T}\}$ 
9:     if there exists a  $x \in S$  such that  $T \cup p_{x,T}$  does not
       generate a new branch node then
10:       $T \leftarrow T \cup p_{x,T}$ 
11:   else
12:     Choose a  $x \in S$  and  $T \leftarrow T \cup p_{x,T}$ 
13:      $K \leftarrow K - \{x\}$ 
14:  $A(T) \leftarrow c(T) + b(T)w$  //The weight of the tree  $T$ 
15:
16: //Branch Optimization Phase 1) Deletion Step
17: Obtain an order  $\sigma$  which sorts the branch nodes in the
    ascending order of the degree in  $T$ 
18: for  $v_d \in \sigma$  do
19:    $T' \leftarrow T - \{v_d\}$ 
20:   for neighbor branch node or neighbor terminal node
      $v$  of  $v_d$  do
21:     Reroute the  $v$ 's closest branch node  $u$  in another
       connected component via its shortest path  $p_{v,u}$ 
22:      $T' \leftarrow T' \cup p_{v,u}$ 
23:     if  $c(T') + b(T')w < A(T)$  then
24:        $T \leftarrow T'$  and  $A(T) \leftarrow c(T') + b(T')w$ 
25:
26: //Branch Optimization Phase 2) Alternation Step
27: Obtain an order  $\sigma$  which sorts the branch nodes in the
    ascending order of the degree in  $T$ 
28: for  $v_a \in \sigma$  do
29:    $T' \leftarrow T - \{v_a\}$ 
30:   Choose a neighbor node  $v_n$  of  $v_a$ 
31:   for neighbor branch node or neighbor terminal node
      $v$  of  $v_a$  do
32:     The shortest path  $p_{v,v_a}$  is replaced by the shortest
       path  $p_{v,v_n}$ 
33:      $T' \leftarrow T' \cup p_{v,v_n}$ 
34:     if  $c(T') + b(T')w < A(T)$  then
35:        $T \leftarrow T'$  and  $A(T) \leftarrow c(T') + b(T')w$ 
36: return  $T$  and  $A(T)$ 

```

V. SIMULATION RESULTS

In this section, we evaluate BAERA in both real networks and massive synthetic networks.

A. Simulation Setup

The simulation is conducted in the following real networks [?]: 1) the Unet network with 49 nodes and 84 links, and 2) the Deltacom network with 113 nodes and 183 links. Many recent SDN works [?], [?] evaluate the proposed approaches

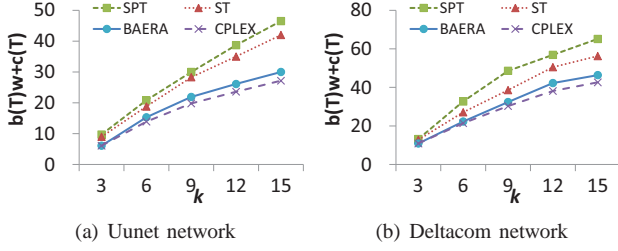


Fig. 3. Varied k in different real networks ($w = 5$)

in real networks with at most hundreds of nodes. By contrast, we also evaluate our algorithm in the networks generated by Inet [?], [?] with tens of thousands of nodes to test the scalability of BAERA. In our simulation, K is chosen randomly from G .

We compare BAERA with the following algorithms: 1) the shortest-path tree algorithm (SPT), 2) a Steiner tree (ST) algorithm [?], and 3) Integer Programming solver CPLEX [?], which finds the optimal solution of the BST problem by solving the Integer Programming formulation in Section III-B. The performance metrics include: 1) the objective value of the BST problem $b(T)w + c(T)$, 2) the number of branch nodes in T , 3) the number of edges in T , and 4) the running time. All algorithms are implemented in an HP DL580 server with four Intel Xeon E7-4870 2.4 GHz CPUs and 128 GB RAM. Each simulation result is averaged over 100 samples.

B. Small Real Networks

In this subsection, we compare the performance of BAERA, ST and SPT with the optimal solutions obtained by CPLEX under different k . Since the BST problem is NP-Hard, CPLEX is able to find the optimal solutions for small instances of the BST problem, and thus we only find the optimal solutions for the Uunet and Deltacom networks. As shown in Fig. 3, the tree T grows and includes more branch nodes as k increases, because a network is inclined to generate a large tree. Nevertheless, BAERA outperforms SPT and ST in the two networks since both the edge number and the branch node number are effectively minimized. In addition, the solutions of BAERA are very close to the optimal solutions.

C. Large Synthetic Networks

In the following, we evaluate BAERA, ST and SPT in large networks with 10000 nodes generated by Inet. Fig. 4(a), Fig. 4(b), and Fig. 4(c) first discover the impact of w with k as 200. Fig. 4(a) demonstrates that the objective value $b(T)w + c(T)$ increases as w grows in all algorithms. For a larger w , BST with BAERA can effectively limit the number of the created branch nodes by slightly increasing more edges necessarily included to span all terminal nodes in K . Nevertheless, BAERA outperforms SPT and ST, especially for a large w , because SPT and ST focus on only the edge number and thus tend to create a tree with more branch nodes. By contrast, the number of branch nodes in the solutions obtained by BAERA is much smaller, but the edge number of BAERA is very close to ST.

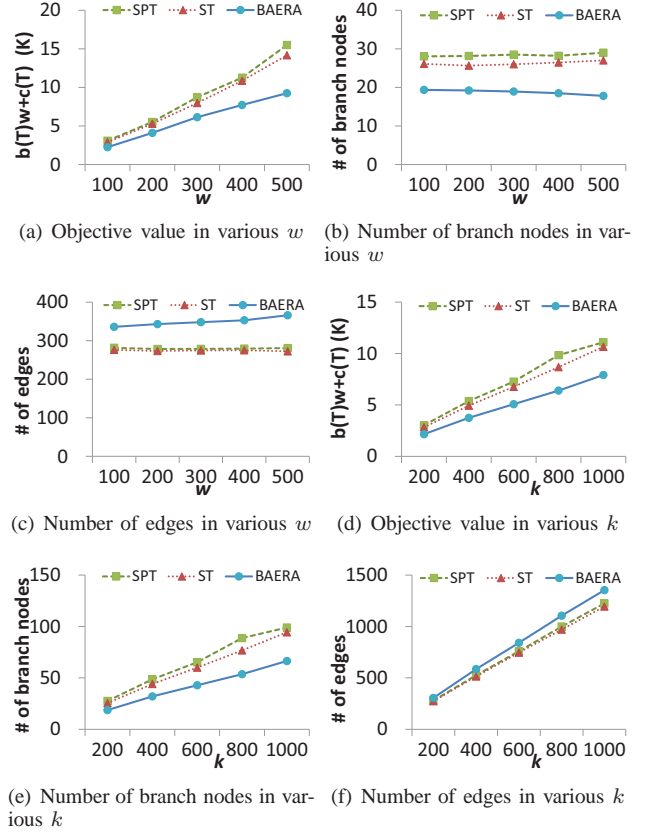


Fig. 4. Varied w and k in the synthetic network by Inet

TABLE I
THE RUNNING TIME OF BAERA IN DIFFERENT k ($|V|=10000$)

k	100	200	300	400
Running time (sec.)	6.064	6.418	6.816	7.430

Fig. 4(d), Fig. 4(e), and Fig. 4(f) evaluate the impact of k with w as 100. As shown in Fig. 4(d), the objective value $b(T)w + c(T)$ becomes larger as k increases, since more branch nodes are necessary to participate in the tree. BAERA still requires fewer branch nodes from Fig. 4(e). Moreover, the increment of the objective value in BAERA grows slower than ST and BT, showing that BAERA can further reduce the total cost in a larger k with the proposed optimization methods.

Table I and Table II evaluate the running time of BAERA with various k and different Inet graph sizes. The running time of BAERA is too small to be measured in the Uunet and Deltacom networks for arbitrary k . It demonstrates that the running time of BAERA only slightly grows for a larger k , and most instances can be solved around 6 seconds when the network has 10000 nodes. In addition, for a smaller graph, ex. 4000 nodes, BAERA takes only 1 second. Therefore, BAERA can both achieve a performance bound (i.e., k -approximation) in theory and find a good solution with small time in practice.

VI. CONCLUSIONS

Traffic engineering and flow table scalability have been studied for unicast traffic in SDN, but those issues in

TABLE II
THE RUNNING TIME OF BAERA IN DIFFERENT GRAPH SIZES ($k = 200$)

$ V $	4000	6000	8000	10000
Running time (sec.)	1.216	2.422	4.148	6.362

multicast SDN have not been carefully addressed. In this paper, therefore, we exploited the branch forwarding technique and proposed Branch-aware Steiner Tree (BST) for SDN. The BST problem is more difficult since it needs to jointly minimize the edge and branch node numbers in a tree, and we proved that this problem is NP-Hard and inapproximable within k . To solve this problem, we designed a k -approximation algorithm, named Branch Aware Edge Reduction Algorithm (BAERA). Simulation results manifest that the trees obtained by BAERA include fewer edges and branch nodes, compared to the shortest-path trees and Steiner trees. In addition, BAERA is efficient to be deployed in SDN because it can generate a scalable and bandwidth-efficient multicast tree in massive networks with only a few seconds.